DOI: 10.14515/monitoring.2022.2.2127

А. В. Резаев, Н. Д. Трегубова

«ЭМОЦИОНАЛЬНЫЙ УТИЛИТАРИЗМ» И ПРЕДЕЛЫ РАЗВИТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Правильная ссылка на статью:
Резаев А. В., Трегубова Н. Д. «Эмоциональный утилитаризм» и пределы развития искусственного интеллекта // Мониторинг общественного мнения: экономические и социальные перемены. 2022. № 2. С. 4—23. https://doi.org/10.14515/monitoring.2022.2.2127.
For citation:
Rezaev A. V., Tregubova N. D. (2022) “Emotional Utilitarianism” and the Frontiers of Artificial Intelligence Evolvement. *Monitoring of Public Opinion: Economic and Social Changes.* No. 2. P. 4–23. https://doi.org/10.14515/monitoring.2022.2.2127. (In Russ.)

DOI: 10.14515/monitoring.2022.2.2127

«ЭМОЦИОНАЛЬНЫЙ УТИЛИТАРИЗМ» И ПРЕДЕЛЫ РАЗВИТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

РЕЗАЕВ Андрей Владимирович — доктор философских наук, профессор, руководитель Международной исследовательской лаборатории ТАНДЕМ, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
E-MAIL: rezaev@hotmail.com
https://orcid.org/0000-0002-4245-0769

ТРЕГУБОВА Наталья Дамировна — кандидат социологических наук, ассистент кафедры сравнительной социологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
E-MAIL: n.tregubova@spbu.ru
https://orcid.org/0000-0003-3259-5566

Аннотация. Вхождение технологий искусственного интеллекта (ИИ) в повседневную жизнь общества ставит перед исследователями новые и нетривиальные задачи, которые могут быть решены только при объединении усилий специалистов из разных областей знания. Авторы статьи предпринимают попытку объединить перспективы социологической теории, философской антропологии и компьютерных наук для ответа на вопросы о том, существует ли альтернатива утилитаризму при разработке ИИ и какие свойства человеческой социальности и эмоциональности могут быть воспроизведены в ИИ, а какие — нет. Основной тезис статьи состоит в том, что концепция индивида, которую предполагает утилитаризм, — весьма спорная в качестве характеристики человека, но неизбежная при описании функционирования искусственного интеллекта. Для обоснования данного тезиса авторы проводят сравнения

“EMOTIONAL UTILITARIANISM” AND THE FRONTIERS OF ARTIFICIAL INTELLIGENCE EVOLVEMENT

Andrey V. REZAEV[1] *— Prof. Dr. habil., Director of International Research Laboratory TANDEM*
E-MAIL: rezaev@hotmail.com
https://orcid.org/0000-0002-4245-0769

Natalia D. TREGUBOVA[1] *— Cand. Sci. (Soc.), Assistant Professor, Chair of Comparative Sociology*
E-MAIL: n.tregubova@spbu.ru
https://orcid.org/0000-0003-3259-5566

[1] St Petersburg State University, St Petersburg, Russia

Abstract. The paper aims to discuss relations between outlooks of utilitarianism on individuals and the subsistence of artificial intelligence (AI) in a society. It deals with two essential questions: a) are there other principles but those developed by utilitarianism for retaining artificial intelligence in everyday life? b) what are the human being attributes that can and cannot be reproduced in artificial intelligence? The starting point for the reflection is the idea that while principles of utilitarianism are debatable when adopted for a concept of the individual in a society, they work fine when appropriated to AI in society. Further, the paper introduces the notion of “emotional utilitarianism”. It characterizes an individual’s emotional dynamics that organize his/her behavior according to the fundamentals of utilitarianism. The authors argue that “emotional utilitarianism” can be reproduced in AI. However, AI fails to execute those features

DOI: 10.14515/monitoring.2022.2.2127

между базовыми принципами социальной и эмоциональной жизни людей и животных, с одной стороны, и агентов ИИ — с другой. Авторы вводят понятие «эмоциональный утилитаризм», которое характеризует эмоциональную динамику индивида, ориентированного на максимизацию эмоциональной энергии. Делается вывод, что «эмоциональный утилитаризм» может быть воспроизведен в искусственном интеллекте — в отличие от собственно человеческих характеристик эмоций и общения. Аргументация статьи базируется на критическом обсуждении и сопоставлении идей ключевых фигур в современной теоретической социологии и философской антропологии: Рэндалла Коллинза, Энн Ролз, Аласдера Макинтайра, Марты Нуссбаум, Мэри Мидгли.

Ключевые слова: искусственный интеллект, искусственная социальность, утилитаризм, общение, эмоции, социальная теория

Благодарность. Исследование выполнено при финансовой поддержке РФФИ и Министерства по науке и технологиям Тайваня в рамках научного проекта № 21-511-52002.

that characterize emotional dimensions of “obschenie” (“social intercourse”) and human interconnectedness. The authors develop their arguments based on the ideas and concepts presented in publications by leading contemporary social theorists — Martha Nussbaum, Alasdair MacIntyre, Mary Midgley, Randall Collins, and Anne Rawls.

Keywords: artificial intelligence, artificial sociality, utilitarianism, social intercourse, emotions, social theory

Acknowledgments. The study was supported by RFBR and MOST, the research project No. 21-511-52002.

Общение и, следовательно, интеллект развиваются только там, где есть долговременные близкие отношения. Если интеллект и может возникнуть в ином контексте, то никто не знает, на что это будет похоже.
Мэри Мидгли [Midgley, 2002: 196]

Исследование принципов работы искусственного интеллекта (ИИ), пределов его развития и следствий проникновения технологий, основанных на ИИ, в повседневную жизнь людей — бесспорно актуальная задача и с практической, и с теоретической точки зрения. Данная задача сегодня стоит перед философами, математиками, представителями технических, естественных и социальных наук и гуманитарного знания.

DOI: 10.14515/monitoring.2022.2.2127

Настоящая статья рассматривает один из аспектов развития новых технологий — а именно, возможности воспроизведения у агентов ИИ человеческих эмоций и социальности. Наша цель — обозначить и обосновать *пределы* развития технологий ИИ в данном направлении. Чтобы достичь данной цели, мы рассматриваем, сопоставляем и критически оцениваем аргументацию ведущих социальных теоретиков и исследователей искусственного интеллекта о базовых свойствах человеческой социальности и возможностях ее имитации агентами ИИ. Статья представляет собой теоретическое исследование, основанное на анализе и синтезе релевантной литературы. В рамках данной работы мы стремимся объединить узкодисциплинарные перспективы; мы анализируем исследования в области теоретической социологии, философской антропологии, моральной философии, философии науки и техники, а также в области компьютерных наук.

Наша аргументация организована вокруг поиска ответов на два вопроса:

— Какие свойства человеческой социальности могут быть воспроизведены в ИИ, какие — нет?

— Есть ли альтернатива утилитаризму при разработке искусственного интеллекта?

Мы полагаем, что эти вопросы связаны: размышление над вторым способно помочь в определении «системы координат» для решения первого вопроса.

Дальнейшее рассуждение организовано следующим образом. Мы начнем с характеристики традиции утилитаризма в моральной философии и соотнесем идеи утилитаризма с принципами работы ИИ. Затем последует сравнение систем предпочтений у людей и агентов ИИ: мы увидим, что функционирование ИИ, в отличие от деятельности человека, хорошо описывается в логике утилитаризма. Рассматривая вопрос об имитации человеческой социальности и эмоций искусственным интеллектом, мы сформулируем концепцию «эмоционального утилитаризма» и критические аргументы в ее адрес, что позволит зафиксировать пределы развития технологий ИИ. Мы также обратимся к этическим вопросам развития ИИ и прокомментируем «тезис об ортогональности». В завершение статьи будут предложены ответы на вопросы, поставленные в ее начале.

Прежде чем перейти к основному рассуждению, следует зафиксировать, *что* мы понимаем под «искусственным интеллектом». «Рабочее определение» ИИ мы формулируем следующим образом: искусственный интеллект представляет собой ансамбль (гармоничную совокупность) разработанных и закодированных человеком рационально-логических, формализованных правил, которые организуют процессы, позволяющие имитировать интеллектуальные структуры, производить и воспроизводить целерациональные действия, а также осуществлять последующее кодирование и принятие инструментальных решений вне зависимости от человека. Современное состояние развития и распространения технологий ИИ характеризуется тем, что агенты ИИ становятся активными посредникам и участниками социальных взаимодействий; данное положение мы фиксируем как возникновение искусственной социальности. Под «агентом ИИ» далее в тексте мы понимаем устройство, деятельность которого опосредует и фиксирует проявления искусственного интеллекта. Детальное обоснование и обсуждение данных определений в соотношении с другими традициями исследования ИИ см в [Резаев, Трегубова, 2019].

DOI: 10.14515/monitoring.2022.2.2127

Традиция утилитаризма в современной моральной философии

Существует ли альтернатива утилитаризму при разработке ИИ? Данный вопрос с необходимостью предполагает предварительное разъяснение того, что есть утилитаризм.

Традиция утилитаризма — одна из наиболее значимых для современной моральной философии. Утилитаризм «в его простейшей формулировке утверждает, что морально правильным действием или мерой государственной политики является то, которое создает наибольшее счастье для наибольшего количества членов общества» [Кимлика, 2012: 27]. Джон Ролз, один из наиболее выдающихся моральных философов современности, утверждал, что утилитаризм — это неявный фон, на котором другие теории должны утверждать и защищать себя [там же].

Проблема соотношения утилитаризма с другими системами является ключевой для современной моральной и политической философии и неоднократно обсуждалась в специальной литературе (см., например, [Scarre, 1996; Griffin, 1991; Резаев, Жихаревич, Трегубова, 2014; Юдин, 2018]). В рамках настоящей статьи, однако, нас интересует другая сторона утилитаризма. Далее мы будем рассматривать не этическую аргументацию утилитаризма, а его «антропологию» — идеи и допущения о человеческой природе, о человеческих желаниях и действиях, на которых основывается эта моральная философия. Представления утилитаристской традиции о природе человека существуют и вне контекста дебатов об этике. Они, прямо или косвенно, лежат в основе утверждений о человеческий действиях и предпочтениях в повседневной жизни и в научной дискуссии (наиболее очевидный пример — бихевиористская исследовательская программа в социальных науках).

Камень преткновения антропологии утилитаризма — это квантификация счастья, количественное сопоставление удовольствий и страданий. Утилитаристы приписывают ценность удовольствию или счастью, предполагая, что эти состояния представляют ценность для тех, кто их испытывает. При этом удовольствие и неудовольствие (счастье и несчастье) поддаются сравнению по количественной шкале «больше-меньше» и для одного, и для многих людей: сегодня я счастливее, чем вчера, а несчастье граждан моей родины сильнее, чем несчастье их соседей. Против такого представления о природе человека, однако, можно выдвинуть как минимум два аргумента.

Первый критический аргумент состоит в том, что индивидуалистическая форма утилитаризма противоречит весьма распространенной концепции ценности. Согласно данной концепции, в мире существует множество вещей, обладающих внутренней ценностью для любой человеческой личности или даже для любого отдельного существа: множество различных форм индивидуальности и способов жизни, распределительная справедливость, красивые ландшафты и т. д. [Резаев, Жихаревич, Трегубова, 2014].

Второй аргумент, развиваемый в работах Аласдера Макинтайра, заключается в том, что «понятие человеческого счастья не является унитарным, простым понятием и не может обеспечить нас критерием для ключевых актов выбора. Если кто-либо предложит нам, в духе Бентама и Милля, руководствоваться при нашем выборе перспективами нашего будущего наслаждения или счастья, подходящим возражением будет такое: „Но каким наслаждением, каким счастьем

DOI: 10.14515/monitoring.2022.2.2127

нам следует руководствоваться?“ <…> Различные наслаждения, как и различные случаи счастья, по большому счету несоизмеримы, не существует шкалы качества и количества, которая могла бы взвесить их. Следовательно, апелляция к критериям наслаждения не подскажет мне, пить пиво или плавать, а апелляция к счастью не позволит мне сделать выбор между жизнью монаха и жизнью солдата» [Макинтайр, 2000: 90—91]. Иными словами, существуют качественно разнородные виды человеческого счастья (и несчастья), несравнимые напрямую. Более того, выбор между ними и поиск оснований для такого выбора является, по-видимому, ключевым обстоятельством жизни человека.

Утилитаризм как теория искусственного интеллекта

Ирония состоит в том, что, рассматривая внутренние механизмы работы искусственного интеллекта, созданного человеком, мы, по-видимому, не можем избежать суждений в духе утилитаризма. Иными словами, *утилитаризм является спорным в качестве антропологии, но хорош (или неизбежен) в качестве характеристики основных принципов существования агентов ИИ*.

У читателя может возникнуть вопрос, какое отношение утилитаризм в принципе имеет к искусственному интеллекту и почему имеет смысл их сопоставлять. Здесь можно привести как минимум два аргумента. Во-первых, технологии ИИ (как почти все технологии) создаются для достижения некоторых, определяемых людьми целей. ИИ отличается тем, что способен выбирать средства для достижения цели и ставить себе промежуточные цели. Соответственно, рассуждения в терминах целей и средств их достижения объединяют традицию утилитаризма и теорию ИИ. Во-вторых, на тесную связь между утилитаристским мышлением и разработками ИИ указывает техническая терминология: одна из ключевых задач в создании агента ИИ состоит в определении его функции полезности (utility function).

Изначально ИИ создавался для решения конкретных задач — например, выиграть шахматную партию. Тогда логика утилитаризма может быть применена следующим образом: выигрыш является «счастьем», проигрыш — «страданием». Однако многие современные разработки ИИ сложнее: поведение робота-уборщика или беспилотного автомобиля явным образом не ориентируется на достижение конкретной цели. В таком случае ориентация на цель становится частным случаем более общего принципа — функции полезности, которая определяет предпочтительность одних состояний окружающей среды по сравнению с другими, при этом соблюдается свойство транзитивности: если A предпочтительнее B и B предпочтительнее C, то A предпочтительнее C[1]. Стюарт Рассел и Питер Норвиг в классической для компьютерных наук монографии «Искусственный интеллект: современный подход» определяют и обосновывают применение функции полезности следующим образом: «Функция полезности отображает состояние (или последовательность состояний) на вещественное число, которое обозначает соответствующую степень удовлетворенности агента. <…> Во-первых, если имеются конфликтующие цели, такие, что могут быть достигнуты только некоторые из них (например, или скорость, или безопасность), то функция полезности позволяет найти приемле-

[1] Обсуждение функции полезности, ее пользы и неизбежности для ИИ см.: What’s the Use of Utility Functions? // Robert Miles. URL: https://www.youtube.com/watch?v=8AvIErXFoH8 (дата обращения: 19.04.2022).

DOI: 10.14515/monitoring.2022.2.2127

мый компромисс. Во-вторых, если имеется несколько целей, к которым может стремиться агент, но ни одна из них не может быть достигнута со всей определенностью, то функция полезности предоставляет удобный способ взвешенной оценки вероятности успеха с учетом важности целей» [Рассел, Норвиг, 2007: 99]. Далее авторы отмечают, что «любой рациональный агент должен вести себя так, как если бы он обладал функцией полезности, ожидаемое значение которой он пытается максимизировать» [там же].

Неизбежно ли для искусственного интеллекта упорядочение состояний? Мы даем утвердительный ответ на этот вопрос — по двум причинам. Первая причина: ИИ создается для решения функциональных задач (выполнения конкретной работы). Изначально ИИ разрабатывается людьми как нечто, ориентированное на достижение целей, включая варианты с постановкой промежуточных целей, оценкой последствий собственных действий и задействованием других, более сложных интеллектуальных операций. Вторая, более фундаментальная причина заключается в том, что в основании функционирования ИИ лежит использование математического аппарата, а это предполагает квантификацию, упорядочивание, наделение определенных состояний окружающей среды численными значениями, с помощью чего и реализуются фиксированные цели.

Математика представляет собой продукт человеческого разума в социальной среде: ей учатся у других, математики продолжают работу своих предшественников, сотрудничают и спорят с другими людьми. Отличительная черта математики, согласно определению Рэндалла Коллинза, одного из наиболее выдающихся современных теоретических социологов, состоит в следующем: те, кто ей занимается, «сосредоточивают свое внимание на чистых, свободных от конкретного содержания формах человеческих коммуникативных операций: на жестах обозначения единиц как эквивалентных друг другу и составлении из них ряда, на операциях более высокого порядка. <…> Математика имеет дело с универсальным и общим, с теми моделями, которые действительно неопровержимым образом обнаруживаются среди универсальных понятий, поскольку темой математики является чистая общность человеческих коммуникативных операций» [Коллинз, 2002: 1129—1131]. Иными словами, математика универсальна — в той мере, в какой два человека, пересчитывая предметы, согласятся друг с другом. Сложный и разветвленный математический аппарат представляет собой возведение рефлексивных конструкций над этими коммуникативными актами [Коллинз, 2002: 1130]. Именно благодаря характеру коммуникативных операций математика допускает квантификацию. А это, в свою очередь, определяет существенные особенности агентов ИИ, основанных на применении рационально-логических, формализованных правил, которые определяют функцию полезности.

Таким образом, *искусственный интеллект предполагает именно то, что приписывают человеку представители утилитаризма, — возможность однозначным и непротиворечивым образом упорядочить собственные предпочтения*[2].

[2] Проблема определения и упорядочения предпочтений является одной из ключевых для современных утилитаристов. Вопрос о том, сводится ли счастье к реализации предпочтений, или, напротив оценка и упорядочивание предпочтений должны исходить из более общего понимания счастья, также остается дискуссионным в рамках утилитаризма [Scarre, 1996]. Далее мы рассуждаем о предпочтениях, поскольку это позволяет сравнивать поведение людей и машин: не вполне ясно, что было бы аналогом «счастья» для ИИ.

DOI: 10.14515/monitoring.2022.2.2127

Системы предпочтений у людей, животных и ИИ

На данном этапе рассуждений мы сталкиваемся с контраргументом. Разве люди не обладают упорядоченной системой предпочтений? Разве они не используют интеллект именно для выстраивания подобной системы? А если так, есть ли разница между человеком и ИИ? Мы соглашаемся с первой частью контраргумента и отвергаем вторую. Действительно, люди (за исключением некоторых пограничных случаев) обладают упорядоченной системой предпочтений, но это не делает их похожими на агентов ИИ. Чтобы понять, чем они различаются, введем в наше рассуждение третий объект — животных.

В дискуссии об ИИ сравнение людей, животных и искусственного интеллекта не редкость, причем исследователи склонны либо противопоставлять человека — с одной стороны, животных и компьютеры — с другой [Wolfe, 1993], либо подчеркивать сходство всех трех вариаций реализации интеллекта [Boden, 2016]. В рамках настоящей статьи мы поступим по-иному: рассмотрим, в чем заключается сходство между человеком и другими сложно организованными животными по сравнению с искусственным интеллектом.

Обратимся для этого к рассуждениям видного британского философа Мэри Мидгли. Анализируя результаты многочисленных исследований по этологии и социобиологии, Мидгли показывает, что различие между человеком и другими животными состоит в степени рефлексии о собственных приоритетах и в степени осознания затруднений при организации их в последовательную структуру. Однако это различие не столь значительно, как мы привыкли полагать. Мидгли утверждает: человеческая рациональность — это не просто «сообразительность». Быть рациональным — значит обладать «определенной структурой предпочтений, системой приоритетов, основанных на чувстве. Такой тип структуры не уникален для людей, он также обнаруживается у высших животных» [Midgley, 2002: 181]. В развитии интеллекта решающую роль играют эмоциональные конфликты, связанные с затруднениями в определении предпочтений и представляющие собой «столь же серьезную угрозу для жизни, что и голод, и более серьезную, чем нехватка орудий» [Midgley, 2002: 200].

Чем отличается система предпочтений высших животных (включая человека) от системы предпочтений ИИ? Наиболее явное различие состоит в степени их непротиворечивости. Поведение ИИ программируется как последовательное, так что для каждого состояния, как было отмечено, определена его предпочтительность по отношению к любому другому состоянию. Для животных и людей это не так: полная непротиворечивость остается недостижимой. Несмотря на работу интеллекта, в поведении всегда присутствуют конкурирующие мотивы, скрывающие различные предпочтения. «Большинство из нас имеют личность, весьма хорошо интегрированную с одной стороны, со стороны, к которой мы проявляем внимание, и фрагментированную с других сторон, к которым мы менее внимательны» [Midgley, 2002: 189].

Данный аргумент с легкостью встраивается в рассуждение о том, что люди — это недостаточно совершенные машины и нам стоило бы стремиться к тому, чтобы быть логически последовательными и менее эмоциональными. Однако мы, вслед

DOI: 10.14515/monitoring.2022.2.2127

за Мидгли, утверждаем другое: *между системами предпочтений человека и ИИ существует качественное, а не только количественное различие*.

Система приоритетов человека и животных определяется нуждами, потребностями, имеющими собственную, внутренним образом определяемую значимость. Именно поэтому система предпочтений нуждается в постоянной переоценке. Наиболее эффективная реализация потребностей, их совмещение в рамках одной жизни и делает существо рациональным. Рациональность заключается не в осуществлении некоторого алгоритма, а в суждении и действии в ответ на многообразные изменения в окружающем мире [3]. Потребности несводимы друг к другу и потенциально конфликтны: «структура [предпочтений] должна состоять из некоторого числа мотивов, вполне отличных и автономных, но приспособленных к тому, чтобы при нормальном взрослении индивида сочетаться в жизни, удовлетворительной для него в его целостности» [Midgley, 2002: 237]. Мотивы определяются характером взаимодействия существа с окружающей средой; в качестве их природных источников Мидгли называет агрессию, секс, доминирование и заботу о потомстве. Качественная несводимость потребностей составляет биологическую основу качественного разнообразия человеческого счастья и страдания, о которых напоминал Макинтайр, критикуя утилитаризм. Также важно, что выстраивание системы предпочтений у людей и других животных предполагает не просто организацию взаимодействия с окружением, но и установление отношений с другими существами.

Агенты ИИ, в свою очередь, не имеют собственных нужд, помимо реализации внешним образом заданной цели или достижения максимального значения функции полезности. Мидгли замечает: «Компьютеры не рациональны; это глупые вещи. Они не знают, что *имеет значение*; они лишь последовательны» [Midgley, 2002: 201][4]. Их предпочтения совершенным образом упорядочены именно потому, что они могут быть *любыми*.

Здесь возникает вопрос: что будет, если наделить искусственный интеллект системой предпочтений, сходной с человеческой? Решение данной задачи, по-видимому, сталкивается с тем, что ИИ работает с помощью математического аппарата, который предполагает квантификацию. При конфликте предпочтений или одно окажется численно большим и потому более важным, или оба внесут свой вклад в поведение машины, или произойдет сбой в ее работе. Но как возможно их совмещение на качественном уровне? Как возможны суждения самого ИИ о том, как совместить разнородные предпочтения за рамками предписанной системы

[3] Для удовлетворения простых потребностей (голод, жажда, сон) достаточно инстинктов, в которых сами потребности и последовательности действий, направленных на их удовлетворение — своеобразные алгоритмы, — «жестко» сцеплены друг с другом. У высших животных интеллект дополняет инстинкт при удовлетворении более сложных потребностей, когда строгая привязка средств к целям неэффективна (например, при выращивании жизнеспособного потомства у теплокровных животных). Это объясняет различия между общественными насекомыми, которые руководствуются только инстинктами, и общественными млекопитающими и птицами с гораздо более разнообразными и индивидуализированными формами отношений даже внутри одного вида.

[4] Мидгли также заявляет: «Люди, программирующие их [компьютеры] должны быть рациональными, то есть должны быть способны видеть приоритеты среди человеческих потребностей» [Midgley, 2002: 201]. Вопрос в том, в какой степени мы может предугадать действия сложного и развитого ИИ, когда задаем для него благую или «безобидную» цель. Обсуждение данной проблемы см. в: Deadly Truth of General AI? // Computerphile. URL: https://www.youtube.com/watch?v=tcdVC4e6EV4 (дата обращения: 19.04.2022).

DOI: 10.14515/monitoring.2022.2.2127

действий? Наконец, что гарантирует внутреннюю связность предпочтений и действий ИИ? Эти вопросы остаются открытыми[5]. В любом случае «рациональность» такого агента лежит за пределами человеческой (и животной) рациональности.

Перспективы развития ИИ: «эмоциональный утилитаризм»

Итак, системы предпочтений у людей и агентов ИИ имеют качественные различия, причем вторые гораздо лучше описываются в терминах утилитаризма, чем первые. Что из этого следует в отношении того, какие свойства человеческого разума и социальности могут быть воспроизведены в искусственном интеллекте?

На первый взгляд, это означает, что ИИ способен совершать действия, направленные на достижение целей, воплощать инструментальный разум, и не способен на решение собственно «человеческих» задач, связанных с эмоциональностью, воображением, достижением согласия. Но такое заключение противоречит тому, что уже сегодня созданы и успешно функционируют роботы и программы, нацеленные на взаимодействия с людьми, будь то уход за пожилыми или игра с детьми, психотерапия или покер. Многое здесь объясняется усложнением математического аппарата (главным образом — механизмов обучения и самообучения), которые позволяют достигать целей неочевидным для людей образом, а также склонностью людей к «одушевлению» партнеров по взаимодействию. Однако на концептуальном уровне остается нерешенным вопрос о том, что в принципе возможно воспроизвести, что — имитировать, а что остается за границами возможностей ИИ.

С позиций социологической теории очевидно, что человеческое мышление социально, оно становится возможным благодаря способности людей вступать в отношения друг с другом, независимые от их сознания, и развивать различные формы общения[6]. Поэтому исходный вопрос может быть переформулирован: какие свойства человека *как социального существа* — и, как следствие, человеческого разума — могут быть воспроизведены в ИИ, какие — не могут?

Начнем с того, что моделировать *можно*. Для этого обратимся к идеям Рэндалла Коллинза, аргументы которого мы уже привлекли для разъяснения природы математики. Коллинз — автор теории ритуалов взаимодействия (interaction ritual theory) [Collins, 2004], ключевое положение которой состоит в том, что реальность взаи-

[5] Представленный анализ выстраивания предпочтений агентами ИИ основывается на определении искусственного интеллекта, приведенном в начале статьи: ИИ определяется как ансамбль формально-логических правил, определяющих некоторый процесс. В этом отношении способ воплощения агентов ИИ оказывается вторичен по отношению к тем процессам, которые искомые правила определяют. Иными словами, ИИ для нас — это формальная логика, выраженная в форме математических выражений, плюс «железо». Контраргумент к данному тезису мог бы заключаться в следующем: при соединении одного с другим (математики с «железом») возникает нечто, обладающее свойствами, не сводимыми к свойствам соединяемых элементов. Данный контраргумент — классический для исследований сознания в современной философии (достаточно вспомнить дискуссию о «китайской комнате» между Джоном Серлем и его оппонентами). Мы признаем, что такой контраргумент возможен, и ответ на него мог бы быть развит в двух направлениях. Первое — анализ того, как именно в современных агентах ИИ «софт» соединяется с материальным воплощением, как материальное воплощение определяет агентов ИИ — в сравнении с тем, как телесное воплощение определяет человека и животных. Второе направление — исследование того, каким образом нечто, возникшее из соединения элементов, могло бы сочетаться с существованием функцией полезности в ее определяющей роли по отношению к агенту ИИ. Мы оставляем эти сюжеты для будущих исследований, здесь же отметим лишь то, что *существующие* агенты ИИ действуют в соответствии с функцией полезности — и, по-видимому, ни с чем иным.

[6] Выше мы показали, как этот тезис применим к математической деятельности: универсальность математики объясняется характером ее коммуникативных актов.

DOI: 10.14515/monitoring.2022.2.2127

модействия с другими людьми является определяющей для человеческого сознания и поведения. «В центре ритуала взаимодействия — процесс, в ходе которого участники создают общий фокус внимания и вовлекаются в телесные микроритмы и эмоции друг друга. <…> Как следствие ритуалов возникают солидарность, символы и эмоциональная энергия индивида. <…> Взаимное вовлечение участников в общие эмоции и фокус внимания создает разделяемый эмоциональный/когнитивный опыт» [Collins, 2004: 47—48]. С точки зрения индивида, «баланс» успешных и неуспешных (солидарных и несолидарных) взаимодействий выражается в уровне эмоциональной энергии — общем эмоциональном фоне краткосрочных эмоций конкретных взаимодействий. Эмоциональная энергия может быть эмпирически зафиксирована: как физический комфорт при вступлении во взаимодействие, как эмоциональная привлекательность общения определенного рода, как живость воспоминаний и переживаний о прошлых взаимодействиях, как легкость синхронизации с собеседником на уровне реплик и движений.

Теория ритуалов взаимодействия является дискуссионной и задумывалась автором как таковая. Коллинз именует ее «радикальной микросоциологией» и стремится объяснить динамикой ритуалов взаимодействия все, что происходит в жизни людей: от истории философии до истории курения, от половых сношений до стремления к богатству. Автор начинает с общего постулата о важности социальных связей для формирования индивида. Однако то, как именно Коллинз понимает «социальность» и как именно раскрывает данный постулат в своей аргументации, отличает его от многих теоретиков социального взаимодействия.

Согласно Коллинзу, поведение людей в самом общем виде описывается принципом максимизации эмоциональной энергии: «Люди не очень хорошо умеют просчитывать издержки и выгоды, но они чувствуют, как двигаться по направлению к целям, потому что на подсознательном уровне могут судить обо всем происходящем в соответствии с его вкладом в фундаментальный мотив — в поиск максимальной эмоциональной энергии в ходе ритуалов взаимодействия» [ibid.: xiii]. Стремление к достижению материальной выгоды (и иных социальных благ) также объясняется социальными (в коллинзовском смысле) мотивами.

Очевидно, что Коллинз использует антропологию утилитаризма, причем антропологию особого рода, которую уместно назвать «эмоциональным утилитаризмом». Согласно Коллинзу, человек — существо социальное, стремящееся к солидарности с другими людьми. Поэтому счастье и несчастье человека поддаются количественной оценке в соответствии с уровнем эмоциональной энергии, который суммирует успехи и неудачи в достижении искомой солидарности. Человеческое мышление объясняется согласно тому же принципу: мы думаем о том, что наиболее эмоционально «заряжено», а стиль мышления зависит от того, у кого мы учимся думать, то есть от групп, в которых взаимодействуем.

«Эмоциональный утилитаризм» Коллинза отличается от «интеллектуального утилитаризма», более привычного для разработок искусственного интеллекта. Какое следствие данный теоретический ход имеет для развития ИИ? Примечательно, что сам Коллинз рассматривает этот вопрос. Он замечает: «по-настоящему „человеческий“ ИИ должен быть способен к настройке на ритмы человеческой речи и их воспроизводству в беседе с реальными людьми; это

DOI: 10.14515/monitoring.2022.2.2127

был бы робот с эмоциональными способностями, который учится использовать символы по-человечески во взаимодействии с людьми, подобно тому как ребенок учится говорить» [ibid.: 182].

Как нам представляется, идеи Коллинза весьма перспективны: они намечают новые горизонты для того, чтобы воспроизводить (и превосходить) в ИИ некоторые человеческие качества. Представим: перед нами два робота, настроенные на ритмическую координацию и «считывание» эмоциональных состояний собеседника (по мимике, жестам, тембру голоса, темпу речи, гормональным изменениям и т. д.). Для первого функция полезности состоит в максимизации координации с собеседником во время взаимодействия, для второго наилучшие состояния собеседника будут предпочтительными состояниями самого робота. В первом случае мы, по-видимому, получим робота — «душу компании», с которым всегда весело; во втором — совершенного альтруиста. Такие варианты ИИ будут обладать некоторыми свойствами человека. Но будут ли они похожи на человека?

Критика «эмоционального утилитаризма»

Вернемся к положению о том, что утилитаризм — это плохая, скажем мягче, — не вполне точная антропология. Из данного положения, если оно верно, следует, что теория ритуалов взаимодействия имеет весьма серьезные ограничения, связанные с пониманием Коллинзом природы социальности и сущности человеческого Я.

Основной контраргумент против «эмоционального утилитаризма» сформулирован в работах Энн Ролз, еще одного крупного социального теоретика. Ролз утверждает, что человеческое общение не сводится к ритуальной динамике, которую характеризует Коллинз, хотя последняя и важна для понимания значительной части содержательной и формальной сторон социальной жизни людей: «Согласно Коллинзу, взаимодействие имеет ценность в основном как источник эмоциональной энергии и культурного капитала. <…> Однако, хотя Коллинз может быть прав в характеристике того, как ритуалы и соответствующие им символы получают свою значимость, он не прав в том, что бóльшая часть взаимодействий являются ритуальными» [Rawls, 1989: 105].

Что во взаимодействии не сводится к стремлению к солидарности, создаваемой в ритуалах? Основной ответ на данный вопрос состоит в указании на связь общения и человеческого Я (Self). Ролз анализирует и реинтерпретирует положения социальной теории Джорджа Герберта Мида и Эмиля Дюркгейма, привлекая внимание читателя к тому, что социальность несводима к поискам солидарности: «Дюркгейм явно указывает на то, что социальной солидарности может быть слишком много (что вызывает альтруистическое самоубийство). Слишком мало солидарности, с другой стороны, приводит к дезинтеграции Я (аномическое самоубийство). Дюркгейм мог бы написать, хотя и не написал об этом прямо, что ценность представляет тот оптимальный баланс солидарности, который и можно назвать „социальностью“. <…> Для Мида <…> социальность — это самоценное (intrinsic) благо, поскольку она служит созданию и поддержанию двух высших самоценных благ [человеческого Я и смысла]. Это обстоятельство не делает социальность „телеологическим“, как заявляет Коллинз, и тем самым чисто утилитаристским

DOI: 10.14515/monitoring.2022.2.2127

благом, потому что она преследует цель не вне, а внутри себя. <…> С точки зрения Мида, нужда в других для поддержания своего Я — это не телеологическая нужда, подлежащая утилитаристскому вычислению, это нужда, внутренне присущая природе Я» [Rawls, 1989: 106].

Иными словами, общение конституирует человеческое Я и тем самым создает субъекта, который затем стремится к взаимодействиям, иногда — к эмоционально интенсивным, солидарным ритуалам взаимодействия. Поиск наиболее интенсивных ритуалов, во-первых, не единственный мотив; во-вторых, разные ритуалы взаимодействия могут быть одинаково привлекательны с точки зрения создаваемой солидарности и баланса эмоциональной энергии. Или, как было отмечено выше, «апелляция к счастью не позволит мне сделать выбор между жизнью монаха и жизнью солдата» [Макинтайр, 2000: 91].

При анализе проблемы создания искусственного интеллекта, подобного человеческому разуму, такая «двуслойная» структура человеческой социальности представляется чрезвычайно важной. На поверхности мы имеем дело с динамикой ритуалов взаимодействия, обусловленной поиском солидарности и определяющей значительную часть человеческого поведения и мышления. Ритуальная динамика поддается частичной или полной квантификации и потому может быть воспроизведена в ИИ. Однако за поверхностью скрывается «ядро» человеческой социальности, определяющее само существование Я, которое зависит от существования других и от взаимодействия с ними. Здесь «другие» — не просто часть окружающей среды (пусть самая важная), «Я» — не просто обозначение механизма подсчета приобретений и потерь. Детальная аргументация в защиту данного положения требует обращения к классическим текстам: от Джорджа Герберта Мида до Мартина Бубера и Карла Ясперса. Данная задача остается за рамками настоящей статьи. Далее мы рассмотрим одну область, в которой отчетливо проявляются ограничения «эмоционального утилитаризма», — исследование эмоций.

Альтернатива «эмоциональному утилитаризму»: эмоции как ценностные суждения

Проблема возникновения у ИИ эмоциональных структур, сходных с человеческими эмоциями, обсуждается в специальной литературе на протяжении нескольких десятилетий [Sloman, Croucher, 1981; Picard, 2000]. И здесь также принципиален вопрос о том, что может быть воспроизведено или имитировано и что воспроизведено быть не может.

Как ограничения «эмоционального утилитаризма» проявляются в трактовке человеческих эмоций? Для ответа на данный вопрос мы предпримем сравнение двух трактовок человеческих эмоций — концепции эмоциональной энергии Рэндалла Коллинза и концепции эмоций Марты Нуссбаум [Nussbaum, 2001; 2004], одной из наиболее важных фигур в современной философской антропологии.

«Эмоциональная энергия» — ключевое понятие для обоснования «эмоционального утилитаризма» как теории и как исследовательской программы. Согласно Коллинзу, эмоциональная энергия поддается количественной оценке (или, по крайней мере, упорядочиванию): о ней можно сказать «больше» или «меньше». Эмоциональная энергия — долгосрочная эмоция, общее состояние индивида, на которое ока-

DOI: 10.14515/monitoring.2022.2.2127

зывают воздействие отдельные взаимодействия и которое служит для них фоном. То, что мы называем эмоциями в обыденном языке, определяется Коллинзом как колебания эмоциональной энергии в конкретных ситуациях взаимодействия. Эмоциональная энергия побуждает, часто на бессознательном уровне, вступать в одни взаимодействия и избегать других. Эмоциональная энергия «заряжает» некоторые объекты (в том числе других людей), связанные с успешными или провальными взаимодействиями в прошлом, придавая им ценность, положительную или отрицательную.

В свою очередь, Нуссбаум характеризует эмоции как «перевороты в мысли» (upheavals of thought). Согласно ее определению, «эмоции — это оценки (appraisals) или ценностные суждения, которые приписывают вещам и людям (persons), находящимся за пределами контроля человека, значимость в отношении благополучия (flourishing) самого человека. Оно [определение] содержит, таким образом, три важные идеи: идею *когнитивной оценки*; идею собственного *благополучия* или *важных целей и проектов*; идею *значимости внешних объектов как элементов собственной системы целей*» [Nussbaum, 2001: 4]. Эмоции предполагают Я и мир объектов, обладающих значимостью, и включают в себя историю отношений с объектами. Кроме того, согласно Нуссбаум, эмоции свойственны не только людям: высшие животные также переживают эмоциональные состояния.

Как соотносятся между собой данная концепция эмоций и идеи теоретиков, которых мы рассматривали выше? При сопоставлении аргументов Марты Нуссбаум и Энн Ролз становится очевидным, что они дополняют друг друга. Для обеих важнейшим элементом «ядра» человеческой социальности является необходимость выделения из окружающего мира значимых других и отношения с ними, выстраивание собственной системы целей с возможностью ее переоценки, а также существование качественно различных эмоциональных состояний, воплощающих оценочные суждения об отношениях Я со значимыми другими.

Характеристика эмоций, которую предлагает Нуссбаум, также хорошо соотносится с анализом систем предпочтений людей и животных Мэри Мидгли. Наличие индивидуальной системы предпочтений (целей и проектов) позволяет говорить о наличии эмоций не только у людей, но и у высших животных. Эмоции не сводятся к сиюминутному удовольствию или страданию, эмоциональная реакция на события определяется их значением для системы (scheme) целей, и ее «пересмотр изменит объекты и конкретные проявления эмоций» [ibid.: 132]. Нуссбаум, как и Мидгли, подчеркивает амбивалентность эмоций по отношению к своим объектам, связанную с противоречивостью системы целей: «мы часто ценим вещи, не спрашивая, согласуются ли наши цели друг с другом; иногда они плохо согласуются, из чего проистекают болезненные эмоциональные конфликты» [ibid.: 49]. Автор дополняет биологическое объяснение амбивалентности эмоций психологическим, указывая на базовый конфликт: объекты эмоций необходимы для собственного благополучия, но не вполне поддаются контролю.

Возвращаясь к сравнению, следует отметить, что между концепциями эмоций Коллинза и Нуссбаум существуют примечательные сходства и не менее важные различия. Сходства суммируются следующим образом. Во-первых, в окружающем мире существуют объекты, обладающие ценностью для индивида. Во-вторых, их

DOI: 10.14515/monitoring.2022.2.2127

ценность определяется историей взаимодействия с объектами того, кто испытывает эмоции. В-третьих, эмоции представляют собой реакцию на ценность объектов и направляют деятельность субъекта, часто — помимо сознательных расчетов. Наконец, эмоции — это долгосрочные явления, которые не сводятся к сиюминутным проявлениям, к краткосрочным реакциям на события.

Различия между рассматриваемыми концепциями заключаются в статусе объектов и в том, что придает им ценность. Для Коллинза ценность объектов — производная от характеристик ритуалов взаимодействия, от динамики солидарности, которая отражается в колебаниях эмоциональной энергии. Для Нуссбаум ценность объектов связана с отношением к личной системе целей и проектов, определяющей благополучие персоны; эмоции «содержат неустранимую отсылку к своему Я (self)» [ibid.: 52]. Для Коллинза объекты эмоций представляют собой структурные части окружающего мира, необходимые элементы ритуалов взаимодействия; они являются внешним и отчасти случайным условием для реализации успешного ритуала взаимодействия (одни люди и символы могут быть заменены другими). Для Нуссбаум объекты отделены от окружения и обладают собственным существованием — в той же мере, что и субъект эмоций. Для Коллинза эмоциональная жизнь человека упорядочена и, потенциально, исчислима; поэтому подлинные конфликты в ней, по-видимому, невозможны. Для Нуссбаум эмоции — это качественно различные состояния (одна привязанность не отменяет и не преодолевает другую); кроме того, эмоции, особенно наиболее важные для нас, амбивалентны по отношению к своим объектам. Для Коллинза эмоции определяют познавательные процессы, однако сами составляют особую предкогнитивную сферу, направляющую деятельность разума и в некотором смысле главенствующую над ней. Для Нуссбаум эмоции — это когнитивные суждения особого рода, связанные с оценкой значимости объектов, они вступают во взаимодействие с другими суждениями и являются частью познавательной сферы человека. Наконец, концепция Нуссбаум предполагает гораздо больше возможностей для «переоценки ценностей» и рефлексии субъекта над собственной эмоциональной жизнью, над собственными целями и планами, в то время как концепция Коллинза, доведенная до логического завершения, подразумевает «эмоциональный детерминизм» при определении системы предпочтений.

Какая из двух концепций эмоций более точная? Обе обладают большой эвристической ценностью, однако мы полагаем, что концепция Нуссбаум более предпочтительна — именно как теория *эмоций*. Динамика эмоциональной энергии в упрощенном виде показывает, как складываются человеческие привязанности, как формируются общности и соответствующие им вкусы и стили мышления. Однако концепция Коллинза не позволяет увидеть и понять важнейшие свойства человеческих эмоций: их качественную несводимость друг к другу, чреватую конфликтами; их амбивалентность; их связь с нашими представлениями о благой жизни, а не только с удовольствием; их когнитивную природу; наконец — существование «созерцательных» эмоций, не связанных с социальными взаимодействиями[7].

[7] Нуссбаум пишет о созерцательной эмоции удивления (wonder), которая позволяет «переместить удаленные объекты в круг системы целей человека» [Nussbaum, 2001: 54], будь то другие люди, явления природы, произведения искусства или нечто иное.

DOI: 10.14515/monitoring.2022.2.2127

Выход на этические проблемы: «тезис об ортогональности»

В завершение настоящей статьи попробуем приложить наши рассуждения к «проблеме ортогональности» — одной из важнейших проблем в осмыслении возможностей и рисков развития ИИ. Современный философ Ник Бостром формулирует «тезис об ортогональности» следующим образом: «Интеллект и конечные цели (final goals) представляют собой перпендикулярные (orthogonal) оси, по отношению к которым возможные акторы могут располагаться где угодно. Иными словами, почти любой уровень интеллекта совместим с почти любыми конечными целями» [Bostrom, 2012: 73].

«Тезис об ортогональности» является возражением на следующее, весьма распространенное суждение: по-настоящему «умный» искусственный интеллект научится понимать, что такое хорошие (с точки зрения человека) цели, и не сможет достигать «глупых» или «злых» целей. Возражение Бострома заключается в том, что уровень интеллекта, понимаемый как способность инструментальной реализации заданных целей с помощью оптимальных средств, и постановка самих целей являются независимыми (ортогональными) свойствами для различных акторов. Под «акторами» (agents) автор понимает любые возможные формы разума, включая те, что нам уже известны (животные, человек, ИИ). Так, очень недалекий человек может ставить благие цели, а очень искусный компьютер — уничтожить вселенную, чтобы вычислить число «пи» с наибольшей точностью.

«Тезис об ортогональности» основан на разделении между целями и средствами, внешними по отношению к целям, которое характерно для утилитаризма. Поэтому, следуя логике настоящей статьи, уместно предположить, что данный тезис справедлив в отношении ИИ (для характеристики которого неизбежны суждения в духе утилитаризма), однако является спорным в отношении человека (ибо спорна сама антропология утилитаризма). Чтобы показать, в чем состоит дискуссионность «тезиса об ортогональности», вернемся к идеям Аласдера Макинтайра.

Бостром, обосновывая тезис, ссылается на различение между существованием и долженствованием: еще Давид Юм писал, что из «есть» не выводится «должно быть»; соответственно, цель не определяется существующими средствами. Это кажется общим местом, однако такое понимание характерно для современного мышления. Альтернативой является аристотелевская традиция добродетелей, реконструируемая Макинтайром: «В рамках классической традиции „человек“ — это „хороший человек“, а „часы“ — это „хорошие часы“ и „фермер“ — это „хороший фермер“» [Макинтайр, 2000: 87—88]. То есть из сущего можно вывести должное: часы, поскольку они существуют, должны показывать время точно, фермер должен усердно трудиться, а человек в принципе — быть добродетельным. Здесь цели и средства не могут рассматриваться по отдельности: мудрость как раз и заключается в том, чтобы выбирать адекватные средства для хороших целей. Связано это с отношением между целями и средствами — не внешнем, как в утилитаризме, но внутреннем (как при общении и поддержании границ своего Я): «Средства и цели могут быть адекватно охарактеризованы без ссылки друг на друга; и для достижения одной и той же цели могут быть использованы совершенно разные средства. Но проявление добродетелей не является в этом смысле средствами достижения человеческого блага. Потому что человеческое благо представлено

DOI: 10.14515/monitoring.2022.2.2127

человеческой жизнью в ее лучших достижениях, и проявление добродетелей есть необходимая и центральная часть такой жизни, а не просто подготовительное упражнение для обеспечения такой жизни» [Макинтайр, 2000: 205][8].

Таким образом, при рассмотрении человеческой жизни и человеческого блага существует альтернатива «тезису об ортогональности», связанная с антиутилитаристским пониманием связи между целями и средствами благой жизни. Однако в отношении искусственного интеллекта такая альтернатива невозможна, если мы принимаем вывод о неизбежности утилитаризма для характеристики принципов действия ИИ. Наше дополнение к тезису состоит в том, что понимание «инструментальной рациональности» ИИ, «ортогональной» конечным целям, следует расширить, включив в нее реализацию функции полезности в соответствии с принципами «эмоционального утилитаризма». Это важно в ситуации искусственной социальности, когда эффективная работа агентов ИИ все больше зависит от того, как они взаимодействуют с людьми [От искусственного интеллекта…, 2020].

Выводы

Мы начинали статью с двух вопросов:

— Какие свойства человеческой социальности могут быть воспроизведены в ИИ, какие — нет?

— Есть ли альтернатива утилитаризму при разработке искусственного интеллекта?

На второй вопрос мы отвечаем отрицательно: деятельность существующих агентов ИИ и их системы предпочтений хорошо описываются в логике утилитаризма, и нет оснований предполагать, что в будущем нас ждет нечто иное.

На первый вопрос можно ответить так: успехи и провалы ИИ в воспроизведении человеческой социальности связаны с тем, насколько хорошо общение и эмоции могут быть описаны в терминах утилитаризма. Для обозначения «пределов роста» ИИ в данной сфере мы вводим понятие «эмоциональный утилитаризм».

Подводя итоги, сформулируем тезис, суммирующий критику «эмоционального утилитаризма». *Познавательная деятельность человека с необходимостью включает в себя эмоциональную сферу с неустранимой отсылкой к своему Я (Self). Эмоциональная сфера содержит не вполне согласованную систему предпочтений, связанную с благополучием субъекта, в которую включается оценка значимых других как отдельных субъектов, а не как элементов окружающего мира.* Если данный тезис верен, то задача воспроизведения подлинно человеческого способа мышления и взаимодействия у искусственного интеллекта представляется недостижимой. Вместе с тем воспроизведение *некоторых* эмоциональных структур, в частности тех, что описываются моделью «эмоционального утилитаризма», возможно — по крайней мере, в теории.

Какие выводы следуют из настоящего рассуждения?

[8] Макинтайровская трактовка добродетелей предполагает ведение человеком социальной жизни в сообществах особого рода. Таким образом, он связывает благую человеческую жизнь с особой социальностью, которая является по отношению к ней средством, неразрывно связанным с целью. В этом отношении сравнение идей Аласдера Макинтайра, Марты Нуссбаум и Энн Ролз могло бы стать предметом для отдельного исследования, которое, однако, выходит за рамки настоящей статьи.

DOI: 10.14515/monitoring.2022.2.2127

Во-первых, *«плохие» (не вполне точные) идеи о человеке оказываются хорошими применительно к искусственному интеллекту*. В попытках понять себя (или убедить других, что мы себя понимаем), люди сформулировали идеи и принципы, позволяющие понять создание собственного разума — искусственный интеллект. Настоящая статья обосновывает, что такие идеи развиваются в рамках утилитаристской традиции, включая то, что мы обозначили как «эмоциональный утилитаризм».

Во-вторых, наш аргумент подтверждает тривиальный, но часто упускаемый из виду факт: *человек и искусственный интеллект представляют собой разные сущности, и, как следствие, взаимодействие между человеком и ИИ отлично от взаимодействия между людьми*. В этом отношении представленное рассуждение развивает экзистенциально-феноменологическую критику искусственного интеллекта, основания которой были заложены Хьюбертом Дрейфусом [Дрейфус, 1978]. В настоящее время в исследованиях ИИ (особенно в узкоспециальных) наблюдаются две тенденции. Первая, техническая, состоит в том, чтобы рассматривать человека как «плохую машину»: с противоречивой «функцией полезности», запутанным «интерфейсом», подверженную внешним влияниям и т. д. Вторая, гуманитарная, заключается в «одушевлении» машины, когда проблемы сводятся к знанию о том, как именно пользователи очеловечивают ИИ — или как это следовало бы делать. Обе тенденции маскируют действительную проблему: мы имеем дело с существами разной природы, с разными принципами деятельности, и осмысление взаимодействий между ними — новая, концептуально захватывающая и прагматически важная исследовательская задача.

Наконец, следует вернуться к тому, что имели в виду основоположники утилитаризма и чего они желали. Джон Стюарт Милль верил в свободу одновременно и как во благо в себе, и как в средство для достижения счастья. Он настаивал, что свой собственный способ «развертывания своего существования» — наилучший именно потому, что он свой, уникальный. Данное стремление считаться с «особенностью индивидов» сталкивается в рамках утилитаризма с серьезными проблемами [Резаев, Жихаревич, Трегубова, 2014]; возможно, другие авторы и другие философские традиции успешнее отстаивали дух миллевских сочинений, чем сама традиция утилитаризма. В этом отношении буквальное воплощение принципов утилитаризма — будь то в человеческой жизни или в работе искусственного интеллекта — вызывает большие сомнения не только в качестве цели, но и в качестве средства.

Список литературы (References)

Дрейфус Х. Чего не могут вычислительные машины? Критика искусственного разума. М.: Прогресс, 1978.
Dreyfus H. (1978) What Computers Can't Do: A Critique of Artificial Reason. Moscow: Progress. (In Russ.)

Кимлика У. Современная политическая философия: введение. М.: Изд. дом Гос. ун-та — Высшей школы экономики, 2010.
Kymlicka W. (2010) Contemporary Political Philosophy: An Introduction. Moscow: Higher School of Economics Press. (In Russ.)

DOI: 10.14515/monitoring.2022.2.2127

Коллинз Р. Социология философий. Глобальная теория интеллектуального изменения. Новосибирск : Сибирский хронограф, 2002.
Collins R. (2002) The Sociology of Philosophies: A Global Theory of Intellectual Change. Novosibirsk: Sibirskii khronograf. (In Russ.)

Макинтайр А. После добродетели: Исследования теории. М. : Академический Проект; Екатеринбург: Деловая книга, 2000.
MacIntyre A. (2000) After Virtue: A Study in Moral Theory. Moscow: Academic Project, Ekaterinburg, Delovaya kniga. (In Russ.)

От искусственного интеллекта к искусственной социальности: новые исследовательские проблемы современной социальной аналитики / под ред. А. В. Резаева. М. : ВЦИОМ, 2020.
Rezaev A. V. (ed.) (2020) Artificial Intelligence on the Way to Artificial Sociality: New Research Agenda for Social Analytics. Moscow: VCIOM.

Рассел С., Норвиг П. Искусственный интеллект: современный подход, 2-е изд. М. : Вильямс, 2007.
Russel S., Norvig P. (2007) Artificial intelligence: A modern approach. 2nd ed. Moscow: Williams. (In Russ.)

Резаев А. В., Жихаревич Д. М., Трегубова Н. Д. Либерализм против утилитаризма»: проблема «особенности индивидов» в социальной теории // Известия вузов. Серия «Гуманитарные науки». 2014. Т. 5. № 1. С. 31—36.
Rezaev A. V., Zhikharevich D. M., Tregubova N. D. (2014) Liberalism against utilitarianism: the problem of the “distinctive individuals” in social theory. *Izvestiya vuzov. Seriya “Gumanitarnyye nauki”*. Vol. 5. No. 1. P. 31—36. (In Russ.)

Резаев А. В., Трегубова Н. Д. «Искусственный интеллект», «онлайн-культура», «искусственная социальность»: определение понятий // Мониторинг общественного мнения: экономические и социальные перемены. 2019. № 6. С. 35—47. https://doi.org/10.14515/monitoring.2019.6.03.
Rezaev A. V., Tregubova N. D. (2019) Artificial Intelligence, On-Line Culture, Artificial Sociality: Definition of the Terms. *Monitoring of Public Opinion: Economic and Social Changes*. No. 6. P. 35—47. https://doi.org/10.14515/monitoring.2019.6.03. (In Russ.)

Юдин Г. Б. Утилитаризм и коммунитаризм: два подхода к проблеме биотехнологического улучшения человека // Вопросы философии. 2018. № 5. С. 114—124. https://doi.org/10.7868/S0042874418050084.
Yudin G. B. (2018) Utilitarianism and Communitarianism: Two Approaches to Biotechnological Human Enhancement. *Voprosy Filosofii*. No. 5. P. 114—124. https://doi.org/10.7868/S0042874418050084. (In Russ.)

Boden M. (2016) AI: Its Nature and Future. Oxford: Oxford University Press.

Bostrom N. (2012) The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*. Vol. 22. No. 2. P. 71—85. https://doi.org/10.1007/s11023-012-9281-3.

DOI: 10.14515/monitoring.2022.2.2127

Collins R. (2004) Interaction Ritual Chains. Princeton: Princeton University Press.

Griffin J. (1991) Modern Utilitarianism. In: Petit P. (ed.) *Contemporary Political Theory*. New York: Macmillan. P. 73—100.

Midgley M. (2002) Beast and Man. The Roots of Human Nature. London: Routledge.

Nussbaum M. (2001) Upheavals of Thought. The Intelligence of Emotions. Cambridge, UK: Cambridge University Press.

Nussbaum M. (2004) Hiding from Humanity: Disgust, Shame, and the Law. Princeton: Princeton University Press.

Picard R. (2000) Affective Computing. Cambridge, MA: MIT Press.

Rawls A. (1989) Interaction Order or Interaction Ritual: Comment on Collins. *Symbolic Interaction.* Vol. 12. No. 1. P. 103—109. https://doi.org/10.1525/si.1989.12.1.103.

Scarre G. (1996) Utilitarianism. London: Routledge, 1996.

Sloman A., Croucher M. (1981) Why Robots will Have Emotions. Proceedings IJCAI. P. 1—10.

Wolfe A. (1993) The Human Difference: Animals, Computers, and the Necessity of Social Science. Berkley: University of California Press.